arXiv:cond-mat/0205525v1 [cond-mat.stat-mech] 24 May 2002

Lévy scaling: the Diffusion Entropy Analysis applied to DNA sequences

Nicola Scafetta[1,2], Vito Latora[3], and Paolo Grigolini[2,4,5].
[1] *Pratt School EE Dept., Duke University, P.O. Box 90291, Durham, North Carolina 27708*
[1] *Center for Nonlinear Science, University of North Texas, P.O. Box 311427, Denton, Texas 76203-1427*
[2] *Dipartimento di Fisica e Astronomia, Università di Catania, and INFN, Corso Italia 57, 95129 Catania, Italy*
[3] *Dipartimento di Fisica dell'Università di Pisa and INFM, Piazza Torricelli 2, 56127 Pisa, Italy*
[4] *Istituto di Biofisica CNR, Area della Ricerca di Pisa, Via Alfieri 1, San Cataldo 56010 Ghezzano-Pisa, Italy*
(February 1, 2008)

We address the problem of the statistical analysis of a time series generated by complex dynamics with a new method: the Diffusion Entropy Analysis (DEA) (Fractals, **9**, 193 (2001)). This method is based on the evaluation of the Shannon entropy of the diffusion process generated by the time series imagined as a physical source of fluctuations, rather than on the measurement of the variance of this diffusion process, as done with the traditional methods. We compare the DEA to the traditional methods of scaling detection and we prove that the DEA is the only method that always yields the correct scaling value, if the scaling condition applies. Furthermore, DEA detects the real scaling of a time series without requiring any form of de-trending. We show that the joint use of DEA and variance method allows to assess whether a time series is characterized by Lévy or Gauss statistics. We apply the DEA to the study of DNA sequences, and we prove that their large-time scales are characterized by Lévy statistics, regardless of whether they are coding or non-coding sequences. We show that the DEA is a reliable technique and, at the same time, we use it to confirm the validity of the dynamic approach to the DNA sequences, proposed in earlier work.

03.65.Bz,03.67.-a,05.20.-y,05.30.-d

I. INTRODUCTION

The recent progress in experimental techniques of molecular genetic shas made available a wealth of genome data (see, for example, Ref. [1]), and raised the interest for the statistical analysis of DNA sequences. The pioneer papers mainly focused on the controversial issue of whether long-range correlations are a property shared by both coding and non-coding sequences or are only present in non-coding sequences [2–5]. The results of more recent papers [6,7] yield the convincing conclusion that the former condition applies. However, some statistical aspects of the DNA sequences are still obscure, and it is not yet known to what an extent the dynamic approach to DNA sequences proposed by the authors of Ref. [8] is a reliable picture for both coding and non-coding sequences. The later work of Refs. [9] and [10] established a close connection between long-range correlations and the emergence of non-Gaussian statistics, confirmed by Mohanty and Narayana Rao [6]. According to the dynamic approach of Refs. [8,11] this non-Gaussian statistics should be Lévy, but this property has not yet been assessed with compelling evidence. The reason for the confusion affecting this issue is deeper than one can imagine, since it essentially depends on the fact the there exists no reliable method of scaling detection. In fact, all the traditional methods of scaling detection on the market, the Detrended Fluctuation Analysis (DFA) [12], the Standard Deviation Analysis (SDA) [8], and the Wavelets Spectral Analysis (WSA) [7,13], are based on the evaluation of the variance of the process, and therefore yield a scaling that is the correct one only if the process under study is Gaussian.

The main purposes of this paper are:

1) To clarify the meaning of scaling as a form of thermodynamic equilibrium that can be reached after a long time transient, throughout which the conventional techniques of analysis can yield misleading information.

2) To show that a new method, the Diffusion Entropy Analysis (DEA), recently proposed in Ref. [14], is able to yield the correct scaling, even when the observed diffusion process is not Gaussian. We shall show that the departure of the correct scaling, detected by means of the DEA, from the results of the traditional methods, all of them being variance-based methods, is a clear indication of the non-Gaussian character of the process under study.

3) To show the DEA in action by means of an application to the study of DNA sequences. As a remarkable result, we shall show that both coding and non-coding DNA sequences depart from Gaussian statistics and produce Lévy diffusion. This will shed light on some still obscure aspects of the statistical properties of DNA.

II. THE MEANING OF SCALING

The reason for the confusion still present in the issue of the extraction of the long-range statistical properties of DNA sequences (and more in general of any time series: heartbeats, earthquakes, oscillations of markets stocks etc.) is essentially due to the fact the there are no reliable methods of scaling detection. To clarify this crucial

aspect we need to discuss, first, what scaling is all about. Scaling is a property of a probability distribution $p(x,t)$, which formally reads as:

$$p(x,t) = \frac{1}{t^{\delta}} F\left(\frac{x}{t^{\delta}}\right). \quad (1)$$

When we deal with a time series or a generic sequence we need first to construct the probability distribution $p(x,t)$. In order to do so we convert with some method, for instance the one used in this paper, the single sequence into many distinct trajectories. These trajectories start at time $t = 0$ from $x = 0$, and then spread over the x-axis, as a result of their, partial or total, random nature. Thus, rather than observing a single trajectory, we are naturally led to evaluate the probability of observing it. In other words, we rest, with theoretical or computational arguments, on the probability of finding the variable x in the interval $[x, x+dx]$ at time t, denoted by us as $P(x,dx,t)$. The probability density, $p(x,t)$, is defined by $P(x,dx,t)/dx$. The meaning of Eq.(1) is that the process is *stationary*, in spite of the fact that the probability density $p(x,t)$ broadens with time. To stress this aspect, let us focus our attention on the probability densities $p(x,t_1)$ and $p(x,t_2)$, at two distinct times t_1 and t_2, with $t_1 < t_2$. Let us squeeze the abscissa scale of the later distribution, $p(x,t_2)$, by the factor $R \equiv (t_1/t_2)^{\delta} < 1$, and then enhance the intensity of the resulting distribution density by multipliying it by the factor $1/R > 1$. If the property of Eq.(1) holds true, then the resulting distribution density is identical to the former, $p(x,t_1)$. This is equivalent to interpreting the distribution density as a form of equilibrium distribution. This property is deeply related to the foundation itself of statistical mechanics [15]. In fact, in the case where the diffusion trajectory is the superposition of many uncorrelated fluctuations, the resulting diffusion process is predicted by the Central Limit Theorem (CLT) to be a Gaussian probability distribution, a special form of canonical equilibrium, and we can refer ourselves to the transient process necessary for the CLT to work as a kind of transition from dynamics to thermodynamics. In this sense the scaling property of Eq.(1) must be interpreted as a form of thermodynamic equilibrium. Note that in the case of ordinary statistical mechanics, when the CLT applies, we have that $\delta = 1/2$ and $F(y)$ is a Gaussian function of y.

According to the new field of Science of Complexity [16,17], a complex process is expected to yield the property of Eq.(1) with $\delta \neq 1/2$ and (or) $F(y)$ being a form different from the Gaussian one (we shall discuss an example of this non-Gaussian form in later sections). Thus, this raises the question of whether a non canonical equilibrium condition can be generated by sequences reflecting complex dynamics. We should consider three different possibilites:

1) Mandelbrot [17] proposes *Fractional Brownian Motion* (FBM) as a condition exceeding the limits of ordinary statistical mechanics. This corresponds to the scaling condition of Eq.(1) with $\delta \neq 1/2$ while $F(y)$ keeps its Gaussian form.

2) Another possible form of violation, naturally stemming from the Generalized Central Limit Theorem (GCLT) [18], rests on Eq.(1) with $\delta > 1/2$ and *F(y) being a Lévy function*, with the asymptotic property $lim_{y\to\infty} F(y) = const/y^{1+1/\delta}$. This means the occurrence of a disconcerting condition, where the second moment of the distribution is infinite. It is obvious that in practice real time series cannot produce this condition, and that the distribution moments of the observed diffusion process are always finite, being an imperfect realization of the diffusion process with infinite moments.

3) Finally, we should consider also the stretched Gaussians emanating from subdiffusion [19]. Actually, this kind of process is not explicitly examined in this paper. We expect that in this case the standard techniques of scaling dectection might do better than in case 2), since the stretched Gaussians are characterized by finite moments. Therefore, we shall focus our attention on both case 1), where the standard techniques are expected to yield exact results, and on case 2), where the standard techniques are expected to fail.

As we shall show in this paper, all techniques currently adopted to detect scaling are explicitly or implicitly based on the measurement of the second moment of the distribution $p(x,t)$. Thus, the scaling revealed by the ordinary techniques of analysis might depart from the genuine scaling of the process under observation, if this is an imperfect realization of a diffusion process with infinite moments. To stress this crucial aspect we adopt for the scaling parameter δ the symbol H, according to a notation proposed by Mandelbrot to honor Hurst [20] (see also Ref. [21]). Notice that a widely adopted method to express the condition of Eq.(1) is given by

$$x \propto t^{\delta}. \quad (2)$$

This way of expressing the scaling condition is the source of misleading procedures. In fact it is usually assumed that it is equivalent to

$$< x^2(t) >^{1/2} \equiv \int_{-\infty}^{+\infty} x^2 \; p(x,t) dx \propto t^{\delta}. \quad (3)$$

We think that it is much more appropriate to use the following notation

$$< x^2(t) >^{1/2} \propto t^{H}, \quad (4)$$

leaving open the possibility that $H \neq \delta$.

In this paper we show that the *Diffusion Entropy Analysis* (DEA) [14] *is the only technique yielding the correct scaling δ when the observed diffusion process departs from the FBM condition.* In fact all the other techniques, including the Detrended Fluctuation Analysis (DFA) [12], the Standard Deviation Analysis (SDA)

[8], and the Wavelets Spectral Analysis (WSA) [7,13], yield a scaling that would be correct only in the FBM case. This is so because, as we shall see, these techniques rest on variance to evaluate scaling. All these techniques, whose limitations are bypassed by the DEA, are in a sense different versions of the same method, to which we shall refer to as the Variance Method (VM). The departure of the correct scaling, revealed by the DEA, from the results of the VM is consequently a proof of the non-Gaussian character of the process under study.

III. THE DIFFUSION ENTROPY ANALYSIS (DEA)

The Diffusion Entropy Analysis (DEA) is based upon the direct evaluation of the Shannon entropy of the diffusion process. In the continuous-space and continuous-time representation for the probability density $p(x,t)$, the Shannon entropy [22] of the diffusion process reads

$$S(t) = -\int_{-\infty}^{\infty} dx\, p(x,t) \ln[p(x,t)]. \tag{5}$$

To show how the DEA works, let us assume that $p(x,t)$ fits the scaling condition of Eq.(1). Let us plug Eq.(1) into Eq.(5). After a simple algebra, we get:

$$S(\tau) = A + \delta\tau, \tag{6}$$

where

$$A \equiv -\int_{-\infty}^{\infty} dy\, F(y)\, \ln[F(y)] \tag{7}$$

and

$$\tau \equiv \ln(t). \tag{8}$$

Eq.(6) shows that if the diffusion process scales with the parameter δ, the resulting diffusion entropy becomes a linear function of the logarithm of t, with a slope equal to δ. *This makes the slope measurement equivalent to the scaling detection, independently of the form of* $F(y)$.

In the case of ordinary Brownian diffusion, $\delta = 1/2$ and $F(y)$ has the following Gaussian form

$$F(y) = \frac{exp\left(-\frac{y^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}}. \tag{9}$$

Thus Eq.(5) becomes

$$S(t) = \frac{1}{2}\left[1 + \ln(2\pi\sigma^2)\right] + \frac{1}{2}\ln(t). \tag{10}$$

In this case, we have assumed the system to be already in the scaling regime state. More in general, we shall have to address the problem of the transition from the dynamic to the thermodynamic (scaling) regime.

IV. LÉVY WALK

The artificial sequences that we shall use in this paper to show the merits of DEA and the limits of VM rests on a dynamic approach adopted years ago to derive Lévy statistics [11,23]. The importance of this approach to Lévy statistics is due the fact that it makes possible, in principle, to use the same perspective as that adopted in Ref. [24]. Bianucci *et al.* [24] discussed the case of a system of interest interacting with another system called *booster* rather than *thermal bath*, to emphasize that no assumption on its thermodynamic nature was done. The basic aspect of the research project of Ref. [24] was that statistical mechanics, in that case ordinary statistical mechanics, had to be derived from merely dynamic rather than thermodynamic arguments. The same approach can be applied to the derivation of Lévy statistics, with only one significant difference: the phase space of the booster rather than being fully chaotic, as in the case of ordinary statistical mechanics, is weakly chaotic [25]. The phase space consists of chaotic and regular regions, and the booster trajectory tends to sojourn for a long time at the border between chaotic and ordered regions. The waiting time distribution is an inverse power law, and, for simplicity, we assume it to be given by

$$\psi(t) = (\mu - 1)\frac{T^{\mu-1}}{(T+t)^{\mu}}. \tag{11}$$

We make the assumption

$$\mu > 2, \tag{12}$$

which ensures the mean waiting time τ_M to get the finite value

$$\tau_M = \frac{T}{(\mu - 2)}. \tag{13}$$

It is evident from this formula that the parameter T, as well as the power index μ, determine the time duration of the sojourn of the trajectory at the border between chaotic and ordered regions. This inverse power law form, and the resulting stickiness, are naturally generated by the self-similar nature of the borders [25]. We call these crucial subsets of the phase space *fractal borders*.

Now, let us assume that one of the variables of the phase space, called ξ, is the generator of the fluctuations that are collected by the diffusing variable x. Since the fractal borders have a finite size, when the trajectory sticks to one fractal border, the variable ξ gets a value that depends on the trajectory position. Let us make also the assumption that there are only two fractal borders, and that their size compared to that of the whole phase space is so small that the variable ξ gets only two distinct values, denoted by us as W and $-W$. As an example of Hamiltonian model generating velocity fluctuations we have in mind the kicked rotor in the so called

accelerating state [26–28]. The booster trajectory moves erratically in the chaotic sea between the two fractal regions, and after a given time sticks to one of the two fractal regions. After an extended time spent in this fractal region it goes back to the chaotic sea, and after a short diffusion process, it either goes back to the earlier fractal region or it goes to the other one. Due to the power law nature of the waiting time distribution of Eq. (11), the sojourn in the chaotic sea can be ignored. As a result of this dynamic process we shall get a sequence such as $W, W, W, W, ... - W, -W, -W, -W, ...W, W, W, ...$ In this paper we set $W = 1$. This is an example of the time series under discussion in this paper. For simplicity, rather than deriving it running a dynamic system, as the kicked rotor in the accelerating state [26–28], we can directly generate the random sequence $\{\tau_i, \xi_i\}$ in the following way: first the numbers τ_i are randomly drawn from the the distribution of Eq.(11); then the value of ξ_i is established by tossing a coin, and it is assumed that the variable ξ gets the specific value ξ_i for the whole time interval τ_i.

To understand the connection between this kind of sequence and Lévy statistics, we have to use the fluctuation ξ to generate diffusion by means of the following equation of motion:

$$\dot{x}\ (t) = \xi(t)\ . \tag{14}$$

As remarked earlier, ξ is a dichotomous variable, i.e. $\xi = \pm 1$, where 1 is a unit of length. The solution of (14) is given by

$$x(t) = x(0) + \int_0^t dt'\ \xi(t')\ , \tag{15}$$

and our final goal is to evaluate $< x^2(t) >$.

As pointed out by Zaslavsky [25], the condition $\mu > 2$, assumed throghout this paper (see Eq.(12)), ensures the stationary condition, which allows us to properly define $\Phi_\xi(t)$, the normalized correlation function of the fluctuation ξ. This important dynamic property, according to the renewal theory [29], is related to $\psi(t)$ by

$$\Phi_\xi(t) = \frac{1}{\tau_M} \int_t^\infty (t' - t)\psi(t')dt', \tag{16}$$

where τ_M denotes the mean waiting time. Using for $\psi(t)$ the expression of Eq.(11) we obtain

$$\Phi_\xi(t) = \left(\frac{T}{t+T}\right)^{\mu-2}. \tag{17}$$

In this case τ_M is given by Eq.(13). Squaring the expression for $x(t)$ given by Eq.(15) and by using the stationary and dichotomous nature of the fluctuation $\xi(t) = \pm 1$ (the latter yielding $< \xi^2 >= 1$), it is easy to prove that the mean square displacement $< x^2(t) >$ is given by

$$\frac{d}{dt} < x^2(t) >= 2 \int_0^t dt'\ \Phi_\xi(t - t')\ . \tag{18}$$

Finally, by using Eq.(17) we get:

$$\lim_{t\to\infty} < x^2(t) > \propto t^{2H}, \tag{19}$$

with

$$H = \frac{4-\mu}{2} \qquad if\ \mu < 3, \tag{20}$$

and

$$H = \frac{1}{2} \qquad if\ \mu > 3 \tag{21}$$

It is therefore evident that $\mu = 3$ is the border between ordinary and anomalous diffusion. As pointed out in Section II, this result can be trusted only in the Gaussian case.

Let us see why this way of evaluating scaling needs some caution. Thank to the condition of Eq.(12), we can define the number $N = [t/\tau_M]$, where $[y]$ denote the integer part of y. In the case $t >> \tau_M$ the number N becomes virtually identical to the number of random drawings of the numbers τ_i and ξ_i. This is equivalent to drawing the N numbers $\eta_i = \xi_i \tau_i$.

1) In the case where the condition $\mu > 3$ applies, this distribution has a finite second moment. Thus, we can use the Central Limit Theorem (CLT), which yields a Gaussian diffusion, and consequently, $H = 1/2$, which correctly reflects the scaling in this case.

2) In the case $2 < \mu < 3$, the second moment of this distribution is divergent, thereby preventing us from using the CLT. However, in this case we use the Generalized Central Limit Theorem (GCLT) [18]. As shown in Ref. [30], this random extraction of numbers yields a diffusion process, described by the probability distribution $p_L(x, t)$, whose Fourier transform, $\hat{p}_L(k, t)$, reads

$$\hat{p}_L(k, t) = exp(b|k|^{\mu-1}t) \tag{22}$$

with

$$b = W(TW)^{\mu-2} sin[\pi(\mu - 2)/2]\Gamma(3 - \mu). \tag{23}$$

The subscript L stands for Lévy. The numerical simulations support this theoretical expectation [30]. Note that this dynamic approach to Lévy statistics coincides with the Lévy walk [29]. The difference between Lévy walk and Lévy flight is well known. In the case of Lévy flight the random walker makes instantaneously jumps of arbitrary intensity. In the case of Lévy walk, instead, it takes the random walker a time proportional to $|\eta_i|$ to make a jump with this intensity. In the case of Lévy flight, the random walker makes jumps of intensity $|\eta_i|$ at regular time intervals.

We note that the scaling of Eq.(1) derives naturally from the joint use of the assumption $x \propto t^{\delta}$ and norm conservation. It is straightforward to show that within the Fourier representation the norm conservation yields $\hat{p}_L(0,t) = 1$. On the other hand, moving from $|k|$ to $|\kappa| = |k|t^{1/(\mu-1)}$ we obtain the time independent Fourier transform $\hat{p}_{ti}(\kappa) = exp(-b|\kappa|^{\mu-1})$, which fits the normalization condition, and yields the scaling

$$\delta = \frac{1}{\mu - 1} \ , \tag{24}$$

which has to be compared to Eq.(20). It is evident that $H \neq \delta$, in this case.

In this paper, we shall focus our attention on the dynamic condition fitting both the condition of Eq.(12) $\mu > 2$, and the condition

$$\mu < 3. \tag{25}$$

This is in line with the arguments of the dynamic approach to DNA of the earlier work of Refs. [8–11], which proved the DNA sequences to be equivalent to a dynamic process fitting both conditions, ensuring *stationarity*, the former, and *superdiffusion*, the latter, at the same time.

There are two important issues to clarify before proceeding with the next sections. The reader can find a detailed account somewhere else [11,31,32]. However, to make this paper as much selfcontained as possible, we shall shortly outline both of them. The first issue has to do with the time required for the GCLT to apply. The work of Ref. [31] shows that the predictions of the GCLT are realized by the following expression for $p(x,t)$:

$$p(x,t) = K(t)p_T(x,t)\theta(Wt - |x|) + \frac{1}{2}\delta(|x| - Wt)I_p(t). \tag{26}$$

where $p_T(x,t)$ is a distribution that for $t \to \infty$ becomes identical to the Lévy probability distribution of the variable x, namely a function whose Fourier transform coincides with Eq.(22), θ denotes the Heaviside step function and $K(t)$ is a time-dependent factor ensuring the normalization of the distribution $p(x,t)$. This contribution to Eq.(26) is a truncated Lévy distribution, the rationale for it being that no trajectory can travel with velocity of intensity larger than W. The trajectories that at time $t > 0$ are still travelling in the same direction as at time $t = 0$ produce two peaks located at the propagation fronts, $x = Wt$ and $x = -Wt$, and their contribution to $p(x,t)$ is given by the second term on the right hand side of Eq.(26). The number of trajectories that contribute to the peaks is given by the function I_p that has been evaluated in detail by the authors of Ref. [31]. Here it is enough to say that these authors find

$$\lim_{t\to\infty}[I_p(t) - \Phi_\xi(t)] = 0. \tag{27}$$

This means that in the time asymptotic limit the peak intensity becomes identical to the correlation function $\Phi_\xi(t)$ of Eq.(17). On the basis of these arguments they reach the conclusion that in the asymptotic time limit Eq.(26) becomes identical to

$$p(x,t) = p_L(x,t)\theta(Wt - |x|) + \frac{1}{2}\delta(|x| - Wt)\Phi_\xi(t), \tag{28}$$

which coincides with the earlier prediction of Ref. [11]. This conclusion seems to be compatible with the results obtained by using the theory of Continuous Time Random Walk (CTRW) [33], although these authors do no refer explicitly to the correlation function $\Phi_\xi(t)$. For an earlier work based on the CTRW see Ref. [34]

To provide an answer to the first question it is enough to rest on the earlier result of Eq.(28). It takes an infinite time for the GCLT to apply: in fact the intensity of the peaks of the propagation front is proportional to the correlation function of Eq.(27), which is not integrable. During this long transient, as we shall see, the DEA gets closer and closer to the true scaling of Eq.(24), while the distribution second moment, which is finite due to the truncation of the Lévy distribution, yields the fake scaling of Eq.(20).

The second issue is less relevant to the main purpose of this paper. It has to do with another approach to the true scaling, already discussed in Ref. [11]. This has to do with the Hamiltonian derivation of Lévy statistics mentioned in Section II. We study the time evolution of the probability distribution of the diffusion variable x, of the fluctuating variable ξ and of all other variables that might be responsible for the fluctuations of ξ. Then, we make a trace over all the "irrelevant" variables, namely, all the variables but x. The resulting equation of motion is not Markovian, and no ordinary method to make the Markovian approximation can be applied. This is so because the projection method yields a time convoluted diffusion equation with a memory term given by the correlation function $\Phi_\xi(t)$ of Eq.(17), which is not integrable. Consequently a new way to make the Markovian approximation also in this case was invented [11]. It was noticed that this approximation changes the time convoluted diffusion equation into a master equation [35]. To derive from it a result consistent with that of the CTRW used in an earlier work of Zumofen and Klafter [36], and with Lévy statistics as well, the authors of Ref. [35] had to use as a bridge the master equation method of Ref. [37]. This master equation gets the form of a fractional derivative, and, the resulting diffusion process coincides with the predictions of the GCLT, with a diffusion strength b that coincides with that of Eq.(23). It comes to be a surprise, therefore, that the recent work of Ref. [32] proves that the exact solution of the time convoluted diffusion equation yields the same scaling as the VM, namely, the scaling of Eq. (20). This suggests that densities and trajectories might not speak the same language in the case

of non-ordinary statistical mechanics, and it makes much stronger than ever the need for detecting the correct scaling of a time series.

V. THE ALGORITHM

Let us consider a sequence of M numbers

$$\xi_i, \quad i = 1, \ldots, M. \tag{29}$$

The purpose of the DEA is to establish the possible existence of a scaling, either normal or anomalous, in the most efficient way as possible without altering the data with any form of detrending. Here we describe the algorithm adopted in this paper.

Let us select first of all an integer number l, fitting the condition $1 \leq l \leq M$. This integer number will be referred to by us as "time". For any given time l we can find $M - l + 1$ sub-sequences of length l defined by

$$\xi_i^{(s)} \equiv \xi_{i+s}, \quad i = 1, \ldots, l, \tag{30}$$

with $s = 0, \ldots, M - l$. For any of these sub-sequences we build up a diffusion trajectory, s, defined by the position

$$x^{(s)}(l) = \sum_{i=1}^{l} \xi_i^{(s)} = \sum_{i=1}^{l} \xi_{i+s}. \tag{31}$$

Let us imagine this position as that of a Brownian particle that at regular intervals of time has been jumping forward of backward according to the prescription of the corresponding sub-sequence of Eq.(30). This means that the particle before reaching the position that it holds at time l has been making l jumps. The jump made at the i-th step has the intensity $|\xi_i^{(s)}|$ and is forward or backward according to whether the number $\xi_i^{(s)}$ is positive or negative.

We are now ready to evaluate the entropy of this diffusion process. In order to do so we have to partition the x-axis into cells of size $\epsilon(l)$ and to count how many particles are found in the cell i at a given time l. We denote this number by $N_i(l)$. Then we use this number to determine the probability that a particle can be found in the i-th cell at time l, $p_i(l)$, by means of

$$p_i(l) \equiv \frac{N_i(l)}{(M - l + 1)}. \tag{32}$$

The entropy of the diffusion process at time l is:

$$S_d(l) = -\sum_i p_i(l) \ln[p_i(l)]. \tag{33}$$

Note that the subscript d stands for *discrete* and serves the purpose of reminding the reader that the numerical evaluation of the diffusion entropy departs by necessity from the continuous-time and continuous-space picture of Eq.(5) The easiest way to proceed with the choice of the cell size, $\epsilon(l)$, is to assume it independent of l and determined by a suitable fraction of the square root of the variance of the fluctuation ξ_i.

In this paper we study sequences of numbers $\xi_i = +1$ or -1. Because at any step, the jump has the intensity equal to 1, the most reasonable choice of the cell size is given by $\epsilon(l) = 1$. In this way any cell corresponds to a unique position $x(l)$ of the diffusion trajectory defined in (30) and (31). Moreover, $\epsilon(l) = 1$ is the square root of the variance of the random dichotomous fluctuation ξ_i of intensity equal to 1.

Few remarks about the meaning of the integer number l are necessary for the reader to understand the content of the next sections. As said before, l is the length of a window moving all over the available sequence to define distinct trajectories. These trajectories are used to produce diffusion, and consequently we shall often refer to l as time. This should not confuse the reader. The adoption of the term time is suggested by the formal equivalence with the processes of either normal or anomalous diffusion, where walker's jumps occur in time. Here, these jumps occur as we move from a sequence site to the next, and consequently time here has to do with the site positions. Furthermore, we shall be often using for this kind of time the symbol t rather than l. This has to do with the fact that for windows of very large size the integer number l becomes virtually indistinguishable from a continuous number. To emphasize this aspect we shall adopt the symbol t rather than l.

VI. TRANSITION REGIME: RANDOM WALK AND LÉVY WALK

In Section II we have shown that scaling is equivalent to thermodynamic equilibrium with the equilibrium distribution $F(y)$. We refer to the transient process necessary to realize this form of thermodynamic equilibrium from the initial condition with all the trajectories located at $x = 0$, as transition from microscopic dynamics to thermodynamics. Here we illustrate this transition in two different cases, ordinary Brownian motion and Lévy walk. In the former case the transition from microscopic dynamics to thermodynamics can be interpreted as a transition from the discrete to the continuous time representation. In the second case the transition is more extended and can be still perceived after reaching the continuous time regime.

A. The transition regime in the case of Brownian walk

The discrete perspective can be illustrated by using the random walk theory that is expected to apply when our dichotomous signal is completely random. In this specific case, with no correlation, the probability $p_m(l)$, for the random walker to be at position m after l jumps of intensity 1 in either positive or negative direction, is determined by the binomial expression [38]:

$$p_m(l) = \frac{1}{2^l}\binom{l}{\frac{l+m}{2}}\frac{1+(-1)^{l+m}}{2}. \tag{34}$$

and the diffusion entropy reads

$$S_d(l) = -\sum_{m=-l}^{l} p_m(l)\, ln[p_m(l)]. \tag{35}$$

In the continuous time limit we expect Eq.(10) to apply. Fig.1 shows that, after a short initial regime, the discrete diffusion entropy converges to the continuous time prescription (solid line in figure). In the case of Brownian walk we can interpret the transition from microscopic dynamics to thermodynamics as the transition from the binomial formula of Eq.(34) to the Gaussian expression of Eq.(9), with $\sigma = 0.5$.

B. The transition regime in the case of Lévy walk

Here we show how to build a sequence corresponding to the prescription of Section IV. In an earlier work [39] the reader can find the illustration of an algorithm that, using a generator of random numbers of the interval $[0, 1]$, creates the waiting time distribution of Eq.(11). Here we illustrate a slightly different method, generating a distribution of integer times that is exactly, rather than approximately, equivalent to a shifted inverse power law. This serves the purpose of making as fast as possible the transition from microscopic dynamics to thermodynamics, without further delay caused by the time it takes the distribution to become the shifted inverse power law of Eq.(11).

To realize this purpose, first of all we need to generate a series of i integer numbers $L(i)$ according to a probability distribution $p(L)$: these numbers can be interpreted as the lengths of strings of the sequence to build up. Then, for any string, we toss a coin and we fill it entirely with $+1$'s or -1's, according to whether we get head or tail. We assign to the integer numbers L(i) the following inverse power law:

$$p(L) = \frac{C}{(T+L)^{\mu}}, \tag{36}$$

where T and $C = \left(\sum_{L=1}^{\infty}\frac{1}{(T+L)^{\mu}}\right)^{-1}$ are two constants related the one to the other in such a way as to realize the normalization condition without the continuous time assumption behind Eq.(11). It is evident that in the asymptotic limit of very large times the distribution of Eq.(36) becomes equivalent to that of Eq. (11).

To create the distribution of Eq.(36) we proceed as follows. We divide the interval of real numbers [0,1] in infinite sectors. The L-*th* sector, R_L, covers the space

$$R_L \equiv \left[X(L), X(L) + \frac{C}{(T+L)^{\mu}}\right), \tag{37}$$

where

$$X(L) = \begin{cases} 0 & \text{if } L = 1, \\ C\sum_{n=1}^{L-1} 1/(T+n)^{\mu} & \text{if } L > 1. \end{cases} \tag{38}$$

The length of the sector R_L is equal to the probability $p(L)$ given by the Eq.(36). Then, by using a computer, we generate a sequence of rational random numbers $\Upsilon(i)$ uniformly distributed between 0 and 1: if the rational number $\Upsilon(i)$ belongs to the sector R_L, the value L will be assigned to the element $L(i)$ of the sequence of integer numbers. The described algorithm and the uniformity of the sequence of rational random numbers $\Upsilon(i)$ assure that the sequence of integer numbers $L(i)$ is distributed exactly according to the power law given by the equation (36). It is worth to point out that this special method of creating the artificial sequence to analyze by means of the DEA is equivalent to that used by Zumofen and Klafter [36]. Of course, due to the time asymptotic equivalence with the condition discussed in Section II, even in this case the thermodynamic regime is characterized by Lévy statistics and the proper scaling is that of Eq.(24). The diffusion entropy, $S_d(l)$ of Eq.(33), is expected to converge asymptotically to the curve of Eq.(6). For example, if we set $\mu = 2/3$ in Eq.(36), on the basis of Eq.(24) we expect a value $\delta = 2/3$. In principle, using the theoretical approach of Zumofen and Klafter [36] we might also evaluate the value of A using Eq.(7). Since this is not very relevant for the present paper, we skip this issue, and we rest on the numerical simulation to conclude that $A = 1$, thereby reaching the conclusion that the asymptotic time limit is well reproduced by

$$S_d(t) = 1 + \frac{2}{3}\ln(t) \tag{39}$$

From Fig. 2 we see indeed that Eq.(39) fits remarkably well the time asymptotic limit of the numerical curve. We see that this limiting condition is reached after a transient that is significantly larger than that of Fig. 1. A satisfactory discussion of this transient will be presented in Ref. [31].

C. DEA and SDA at work

In this Subsection we show the benefit of the joint use of DEA and SDA. The standard deviation at the diffusion time l, $D(l)$, rests on the following prescription

$$D(l) = \sqrt{\frac{\sum_{n=1}^{M-l}(x_n(l) - \bar{x})^2}{M-l-1}}, \tag{40}$$

where, according to the notation of Section V, M is the sequence length, l denotes the width of moving windows necessary to create distinct trajectories and $\bar{x}$ denotes the mean value of $x(l)$.

According to the theoretical remarks of Section V, the adoption of this method applied to an artificial sequence generated by the inverse power law distribution of Eq.(36), with $2 \leq \mu \leq 3$ should yield the true scaling of Eq.(24). The SDA should generate the Hurst scaling of Eq.(20). We make the analysis of five artificial sequences with the power indices: $\mu = 2.8$, 2.6, 2.5, 2.4, 2.2. We note that at both $\mu = 3$ and $\mu = 2$ the two predictions yield the same values $\delta = H = 0.5$ and $\delta = H = 1$, respectively. Therefore we focus our attention on the intermediate values of μ. For these intermediate values the correct scaling, namely the Lévy scaling, yields $\delta =$ 0.556, 0.625, 0.667, 0.714, 0.833, respectively, while the Hurst scaling is expected to be $H =$ 0.6, 0.7, 0.75, 0.8, 0.9, respectively. For the sake of reader's convenience this situation is summarized in Table I. The numerical results illustrated in Fig. 3 provide a strong support to the theoretical arguments of Section V, and to our claim about the accuracy of the DEA. In fact, we see that the DEA yields a remarkable agreement with the Lévy scaling, while the scaling detected by the SDA virtually coincides with the Hurst scaling.

In general, when the secrete recipe driving the sequence under study is not known, the comparison between the DEA and SDA results plays an important role to assess the statistical nature of the process. In fact, in the case of Lévy statistics, it is easy to show, using Eq.(20), that δ is related to H by

$$\delta = \frac{1}{(3-2H)}. \tag{41}$$

In the FBM case, according to theoretical arguments of Section II, we have

$$\delta = H, \tag{42}$$

and this equality can be considered as a plausible indication that the Gausssian condition applies. The results of Fig.3 fit Eq.(41), thereby confirming the Lévy nature of the diffusion process.

VII. APPLICATIONS TO DNA SEQUENCES

In the last few years, thanks to the recent progress in experimental techniques in molecular genetics, a wealth of genome data has become available (see for example Ref. [1]). This has triggered a large interest both in the study of mechanics of folding [40], and on the statistical properties of DNA sequences. In particular, genomes can be considered as long messages written in a four-letter alphabet, in which we have to search for information (signal). Recently, there have been many papers pointing out that DNA sequences are characterized by long-range correlation, this being more clearly displayed by non-coding than by coding sequences [2,5,11,12].

In this section we will study a large sample of DNA sequences (a dozen of both coding and non-coding sequences). In particular we discuss in detail three DNA sequances:
– the human T-cell receptor alpha/delta locus (Gen Bank name HUMTCRADCV) [12], a *non-coding* chromosomal fragment (it contains less than 10% coding regions);
– the Escherichia Coli K12 (Gen Bank name ECO110K) [12], and the Escherichia Coli (Gen Bank ECOTSF) [10], two genomic fragments containing mostly *coding regions* (more than 80% for ECO110K).
The three sequences have comparable lengths, $M =$ 97634 basis for HUMTCRADCV, $M = 111401$ basis for ECO110K and $M = 91430$ basis for ECOTSF, respectively. The first two sequences have been analyzed in Ref. [12] by means of the Detrended Fluctuation (DFA). The fundamental difference between them is that the non-coding sequence, namely HUMTCRADCV, shows the presence of long-range correlation at all scales, while the sequence ECO110K, a coding sequence, shows the presence of long-range correlation only at the short-time scale. The third sequence, ECOTS, has been studied in Ref. [10] with the interesting conclusion that the large-time scale shows non-Gaussian statistics. The authors of Ref. [12], using the illuminating example of the lambda phage genome, pointed out that the DFA does not mistake the presence of patches of different strand bias for correlation. This is an important property, shared by the DEA method, which is widely independent of the presence of biases, since the entropy increases mainly as a consequence of the trajectories departing from one another. In this Section we show that *the DEA method makes it possible to relate the non-Gaussian statistics and the anomalous scaling of the large-time scale to the same cause: the onset of Lévy statistics* .

A. The numerical representation of DNA

The usual way to study the statistical properties of DNA is to consider a sequence of four bases: adenine,

cytosine, guanine, and thymine (respectively A, C, G, and T), at the simplified level of a dichotomous sequence of two symbols, purine (for A and G) and pyrimidine (for C and T). A trajectory, the so-called DNA walk, can be extracted by considering a one-dimensional walker associated to the nucleotide sequence in the following way: the walker takes one step up when there is a pyrimidine in the nucleotide and a step down when there is a purine. The DNA sequence is therefore transformed in a sequence $\xi_i, \quad i = 1, ..., M$, of numbers $+1$ or -1.

As pointed out at the end of Section V, we associate the site position along the sequence with time. Thus, i is conceived as a discrete time, and the walker makes a step ahead or backward, according to whether at time i the random walker sees $+1$ or -1, namely if the i-th site of the DNA sequence hosts a pyrimidine or a purine. The displacement of the walker after l steps is $x(l) = \sum_{i=1}^{l} \xi_i$ and is reported in Fig. 4 for the three sequences under consideration.

B. The three variance methods (VM) at work: non-coding and coding DNA sequences

This section is devoted to illustrating the three different realizations of the variance method (VM), namely, the Detrended Fluctuation Analysis (DFA) [12], the Standard Deviation Analysis (SDA) [8], and the Wavelets Spectral Analysis (WSA) [7,13]. We have already discussed the first two methods in the previous sections. We have also showed some results of the application of SDA to an artificial sequence in section VI C. As to the WSA, it was first adopted to study DNA sequences by Arneodo and collaborators in Ref. [13], and it consists in reporting the square root of the wavelet variance. In this way, the scaling is comparable to those detected by DFA and SDA, and, as we shall see, it gives indeed the same results.

The first property we notice is that all the three series present "patches", i.e. excess of one type of nucleotide. In the DFA of ref. [12], Stanley and collaborators adopt a detrending procedure to detect the true scaling, since the steady bias hidden in the data can produce effects which might be mistaken for a striking departure from Brownian diffusion, while the interesting form of scalings must be of totally statistical nature. They define a detrended walk by subtracting the local trend from the original DNA walk and then they study the variances $F(l)$ of the detrended walk. If the walk is totally random, as in the ordinary Brownian motion, no correlations exist and $F(l) \sim l^{1/2}$. On the contrary, the detection of $F(l) \sim l^H$ with either $H > 1/2$ or $H < 1/2$ is expected to imply the presence of extended correlation, which, in turn, is interpreted as a signature of the complex nature of the observed process.

To illustrate the results of these authors, let us limit to the long-time region the adoption of the symbol H, which, according to Section II, is used by us to denote the scaling emerging from the VM. When the VM method is applied to the short-time region let us call the scaling parameter determined by the VM with the symbol H'. Stanley *et al.* [12] found a scaling exponent $H' = 0.61$ for the non-coding intron sequence HUMTCRADCV, and $H' = 0.51$ for the intronless sequence ECO110K. They claim that their detrending method is able to avoid the spurious detection of apparent long-range correlations which are the artifacts of the patchiness.

We are now ready to show the three methods at work on the DNA data sets we want to study in this paper.

Non-coding DNA. Fig.5 refers to the sequence HUMTCRADAVC and shows that, within the statistical error, the three VM techniques yield the same long-time scaling, more precisely the three scaling exponents H obtained are 0.59 ± 0.01 (SDA), 0.60 ± 0.01 (DFA), 0.61 ± 0.01 (WSA).

Coding DNA. In Figs. 6 we study the two sequences ECO110K and ECOSTS, and we show that the same equivalence applies to both short-time and long-time scaling. In fact for both sequences we find that H' is 0.53 ± 0.01 (SDA), 0.52 ± 0.01 (DFA), 0.52 ± 0.01 (WSA), and that H is 0.73 ± 0.01 (SDA), 0.75 ± 0.01 (DFA), 0.74 ± 0.01 (WSA).

Before moving to illustrate the results obtained by the DEA, some comments are in order. DFA detects the scaling in the long-time region later because of the detrending that cuts off long local trend. In Ref. [12], Stanley and collaborators were interested in studying the scaling in the short-time region in order to distinguish the non-coding from the coding DNA sequences. The DFA aims at making more visible this regime. However, we think that it is more convenient to study the signal as it is, since detrending might erase important information as well as the deceiving indication of a correlation that does not exists.

Figs. 5 and 6 show that there is no difference between SDA and WSA. This is because the Wavelet Transform behaves like the Fourier Transform that studies the variance of the signal. Therefore, WSA, as Fourier Spectral Analysis, can detect the true scaling only in the Gaussian case. In all other cases, WSA detects only the variance scaling, and this, as pointed out in Section IV, may not coincide with the true scaling.

C. The Copying Mistake Map: a model for DNA sequences

According to the dynamical model of Ref. [8] a *non-coding* DNA sequence corresponds to an artificial sequence with inverse power law long-range correlation as the Lévy walk of Section IV, examined by means of the

DEA in Section VI. On the other side, a *coding sequence* can be obtained by adopting a kind of generalization of the Lévy walk. This generalization becomes a model called Copying Mistake Map (CMM) [8]. This model rests on two sequences of +'s and −'s, running independently the one from the other. The former sequence is the correlated sequence studied in Section VI C by means of the joint use of DEA and SDA. The latter sequence is obtained by tossing a coin. According to the CMM, the generic $i-th$ site of the DNA sequence is assigned the symbol pertaining to the $i-th$ site of either the former or the latter sequence. The former sequence is selected with probability p_L and the latter sequence with probability $p_R = 1 - p_L$. In the case of coding sequences usually the condition

$$p_R \gg p_L \tag{43}$$

applies. The authors of Ref. [10] pointed out that the CMM model is equivalent to an earlier model [5,41] called Generalized Lévy Walk (GLW). The CMM (and the GLW, as well, of course) yields, for short times, a diffusion process indistinguishable from ordinary Brownian motion. At large times, however, the long-range correlation predominates. In Ref. [10] the CMM was adopted to account for the properties of prokaryotes, for which a significant departure from Gaussian statistics occurs. One of the coding sequences studied here, namely ECOTSF, is the same as one discussed in Ref. [10]. It produces strong deviations from Gaussian statistics. On the basis of that, and of the results of Section VI B, we expect also for coding sequences at large times a scaling parameter δ corresponding to the Lévy statistics, and so, to the prediction of both Eq.(24) and Eq.(41).

The CMM is a model flexible enough as to move from the Gaussian to the Lévy condition. This is done simply setting $p_R = 0$. On the other hand, if the condition of Eq.(43) applies, in the long-time limit we expect the condition of Lévy statistics will emerge again. This is so because the most evident sign of Lévy statistics is given by the power law character of the distribution tails. The correlated component of the CMM model results in a process of diffusion faster than ordinary diffusion, and so faster than the diffusion generated by the random component. As a consequence, the distribution tails are forced to get the character of an inverse power law.

D. DEA at work: non-coding and coding DNA sequences

By using the DEA algorithm we can detect the existence of scaling, either normal or anomalous, Gaussian or Lévy, in a very efficient way, and without altering the data with any form of detrending. We analyze the data of both the coding and non-coding sequences. Starting from the sequence $\xi_i,\ i = 1, ..., N$ we create the diffusion trajectories and we compute the diffusion entropy $S_d(l)$ according to equation (33). The results are reported in Figs. 7-9. We determine the scaling as the slope of the tangent of the curve $S_d(\tau)$. As for the second moment scaling, called H or H', according to whether it refers to long or short times, we adopt for the DEA scaling the corresponding symbols δ and δ'. It is evident that δ is the true scaling. As to the meaning of δ', it will be discussed at the end of this section.

Non-coding DNA. First of all let us consider the *non-coding sequence* HUMTCRADCV. Fig. 7a shows that the DEA results in what seems to be a time dependent scaling. This is pointed out by means of the two straight lines of different slopes, $\delta' = 0.615 \pm 0.01$ and $\delta = 0.565 \pm 0.01$ Anomalous diffusion shows up at both the short-time and the long-time scale, and this seems to be a common characteristic of non-coding sequences, supported also by the application of our technique to other non-coding DNA sequences. Moreover, we notice that the scaling in the short-time regime $\delta' = 0.615 \pm 0.01$ coincides with the value found by means of the DFA analysis [12], $H' = 0.61 \pm 0.01$. The authors of Ref. [12] assign this scaling value to both the short and the long-time regime, while the DEA detects a different scaling at long times. Fig. 7b shows the result of the DEA applied to an artificial sequence built up according to the CMM prescription so as to mimic the sequence HUMTCRADCV. We use both μ and p_R as fitting parameters. In this case, the intensity of the random component is not predominant, as in the case of the coding sequences, which are known [10] to require the condition of Eq.(43). In fact, in this case the best fit between the real and the CMM sequence is obtained by setting $p_R = 0.56 \pm 0.02$. As to μ, the value of it emerging from this fitting procedure, is considered by us to be the best estimate of this inverse power law index. This value is $\mu = 2.77 \pm 0.02$. If we plug it into Eq.(20), we get $H = 0.615 \pm 0.01$, which is in fact the scaling detected in Ref. [12]. This means that the short-time region obeys the FBM statistics. If we plug it into Eq.(24) we obtain $\delta = 0.565 \pm 0.01$, which is the slope of the DEA curve in the long-time regime, thereby proving that the relation between δ, the true scaling, and H obeys the condition of Eq.(41), which is, as erlier pointed out, a clear indication of Lévy statistics. We consider this to be a compelling evidence that at this long-time scale Lévy rather that FBM diffusion is generated.

Coding DNA. In Figs. 8 and 9 we turn to the more delicate problem of *coding sequences*. The first sequence (ECO110K) has already been studied by means the DFA analysis in Ref. [12]. The DFA finds $H' = 0.52 \pm 0.01$ at the short-time scale and $H = 0.75 \pm 0.01$ in the large-time scale. The second sequence (ECOTSF) has been analyzed in Ref. [8] by using four different methods. The first was the SDA discussed in Section C. This is a method of analysis less sophisticated than the DFA, since does not

imply any local detrending. The second and third methods were the DFA and the Hurst analysis [21], respectively. The fourth method used was the Onsager regression analysis, a method that, in that context, provides information on the correlation function of the fluctuation ξ, which has an inverse power dependence on time l with the power index $\beta = \mu - 2$. The authors of Ref. [8], by using essentially the first method and the Onsager regression analysis, reached the conclusion that the most plausible value of the scaling parameter in the long-time region is $H = 0.75 \pm 0.01$ that is equivalent to the exponent $H = 0.74 \pm 0.01$ found in Figs. 6. It is interesting to remark that the coincidence among the different predictions about scaling, and especially that between the second moment technique and the Hurst analysis, implies the adoption of the Gaussian assumption [42]. On the other hand, when that condition does not apply and the two scaling predictions are different, to the best of our knowledge, it does not seem to be known what is the meaning of any of them. Furthermore, the authors of Ref. [10] pointed out that the statistics of the long-time regime is too poor to support any claim on the departure from the Gaussian condition. In conclusion, in Ref. [10] the claim that the DNA statistics is of Lévy kind was essentially based on the assumption that the dynamical theory of Refs. [8,11] is a reliable approach to the statitstics of DNA sequences. No direct evidence was provided.

The DEA method allows us to prove that the conjecture of the authors of Ref. [10] is correct: the results illustrated in Figs. 8 and 9 afford a convincing proof that the DNA statistics is of Lévy kind. Figs. 8a and 9a clearly show the difference between the slope at short time (respectively $\delta' = 0.52 \pm 0.01$ in Fig. 8a and $\delta' = 0.53 \pm 0.01$ in Fig. 9a) which, in this case, is very close to that of ordinary random walk, and the slope at long time that corresponds to $\delta = 0.665 \pm 0.01$. Since we know that in both cases the long-time slope provided by the DFA is $H = 0.75 \pm 0.01$, we conclude that in both cases the condition of Eq. (41), indicating Lévy statistics, applies. Figs. 8b and 9b aim at fitting the curves produced by the DEA method, applied to the real sequences by means of the CMM model. The purpose is not only that of proving that the CMM can become so close to the real results as to be virtually indistinguishable from them. It is also a way, already applied in Fig.7b, to derive very accurate values for the power index μ. A very good agreement is obtained by setting $p_R = 0.943 \pm 0.01$ for ECO110K (Fig.5b) and $p_R = 0.937 \pm 0.01$ for ECOTSF (Fig.7b). The very good fitting accuracy supports the physical reasons that led the authors of Ref. [8] to propose the CMM model for coding sequences. In fact, with the large weight, $p_R = 0.937 \pm 0.01$, assigned to the random component, the scaling values become $\delta' = 0.52 \pm 0.01$ and $\delta' = 0.53 \pm 0.01$, namely, very close to the conventional scaling $\delta = H = 0.5$. This normal condition lasts for an extended period of time, and eventually, at larger times the transition to a larger scaling takes place.

We note that the authors of Ref. [13] find anomalous diffusion in a statistical condition that they claim to be Gaussian. According to the result of Ref. [11], the Gaussian condition is incompatible with a stationary diffusion process generated by a dichotomous fluctuation yielding a non integrable correlation function with an inverse power law character. This dichotomous fluctuation is expected to generate Lévy rather than Gauss statistics. The authors of Ref. [9] studied under which physical condition FBM is allowed to show up, in apparent conflict with the conclusions of Ref. [11]. With the help of a fractal model for the DNA folding, the authors of Ref. [9] proved that FBM, advocated by the paper of Ref. [13], is possible as a form of non-stationary process. Thus, in principle, the arguments of the work of Ref. [11] would not rule out the possibility that the changing slope is a manifestation of a FBM with a time dependent scaling. This would be another form of transition from dynamics to thermodynamics, of extremely large time duration. However, this way of establishing a compromise between the compelling prediction of the GCLT, according to which a dichotomous process with long-range correlation ($2 < \mu < 2$ must produce Lévy statistics, and the conclusion of some authors that this statistics is Gaussian [13] is ruled out by the statistical analysis of the present paper, which is made much more accurate than the earlier approaches by the DEA. This is made compelling not only by the results illustrated in Figs. 8 and 9, but also by a plenty of statistical measurements on different DNA sequences, reported in Table II. All these results prove that the equality of Eq.(41), implying Lévy statistics, applies to both kind of sequences. This means that in both cases the long-time limit is characterized by Lévy statistics and that this is the form of non-Gaussian statistics revealed by the analysis of Ref. [10].

We can now address the delicate problem of the transition from δ' to δ. On the basis of the results of Figs. 8 and 9, we would be tempted to conclude hat the CMM is a reliable dynamical model for DNA sequences. If this is correct, the transition from δ' to δ is really a time dependent scaling. In fact, according to the CMM the short-time scale is dominated by the random component, due to the fact that the condition of Eq.(43) applies. In the case of Fig. 7 the transition from δ' to δ is probably dominated by a completely different effect. This is the slow transition from dynamics to thermodynamics discussed in Section IV

E. Significance of the results obtained

To properly appreciate the significance of the results of this section, it is necessary to say a few words about the two different scaling prescriptions of Eqs.(24) and

(20). The scaling prescription of Eq.(20) is determined by the adoption of the variance method, as clearly illustrated by the dynamical approach to the DNA sequences of Ref. [8]. This prescription is not ambiguous if the condition of Gaussian statistics applies. In fact, a Gaussian distribution drops quickly to zero, and the existence of a finite propagation front does not produce any significant effect. It has to be pointed out, in fact, that the adoption of the Brownian landscape proposed in the pioneer papers of Refs. [2,5,12] implies the existence of a propagation front moving with ballistic scaling ($\delta = 1$). In other words, if we find a window of length l filled with only 1 's or with only -1's, this means a trajectory travelling with uniform velocity, and the x-space at distances from the origin larger that l is empty. The existence of a propagation front does not have big consequences in the case of Gaussian statistics, since the population at the propagation front is essentially zero in that case. It is not so in the case of Lévy statistics, though, due to the existence of very long tails in that case. Therefore the Lévy processes resulting from these sequences are essentially characterized by the presence of two distinct scaling prescriptions, the Lévy prescription of Eq. (24), concerning the portion of distribution enclosed between the two propagation fronts, and the scaling $\delta = 1$, of the propagation front itself. The scaling of the variance of Eq. (20) does not reflect correctly either of these two different scaling prescriptions, being a kind of compromise between the two. The scaling of the distribution enclosed by the two propagation fronts is, on the contrary, a genuine property that corresponds to the prediction of the GCLT [18]. It is very satisfactory indeed that the DEA method makes this genuine form of scaling emerge. Furthermore, the DEA is a very accurate method of scaling detection, as proved by the fact that it reveals the existence of Lévy statistics in the case of the coding sequence. In this case, as pointed out by the authors of Ref. [8], the ordinary methods become inaccurate due to the poor statistics available in the long-time limit.

Another important result of this section is that it confirms the validity of the CMM model. This model is expected to generate Lévy statistics not only in the case of non-coding sequences, where it is easier to reveal this property. It predicts Lévy statistics also in the case of coding sequences as the one here analyzed. In Ref. [8] the emergence of Lévy statistics was conjectured but not proved, due to the fact that in that paper the observation was made monitoring the probability distribution $p(x,t)$. As already pointed out, the lack of sufficient statistics makes it difficult to assess if the distribution $p(x,t)$ has, or not, tails with an inverse power law character. In Ref. [10] a clear deviation from the Gaussian condition was detected in the long-time limit, but, again, no direct evidence was found that this deviation from Gaussian statistics takes the form of Lévy statistics. The results of this section prove, with the help of the artificial sequences of Section VI B, that the DEA is a method of analysis so accurate as to assess with good accuracy the property of Eq.(41), and with it, the emergence of Lévy statistics for both coding and non-coding sequences.

In conclusion, this paper lends support, with the help of the DEA, an efficient technique of scaling detection, to the claims of Allegrini *et al* [11] about the controversy between Voss [4] and the authors of Ref. [2]. The differences in the findings of the groups-long-range correlations being ubiquitous in DNA sequences by Voss [4] and such correlations being absent in Ref. [2], motivated the authors of Ref. [11] to develop a phenomenological model, the CMM, that might have mitigated the differences between the two apparently conflicting perspectives. The validity of the point of view of these authors is fully confirmed, since Lévy statistics, and consequently long-range correlations, seem to be ubiquitous, being a property of the long-time regime of both coding and non-coding sequences, while the properties of ordinary Brownian motion are confined to the short-time regime of coding sequences.

VIII. CONCLUSIONS

This paper shows that long-range correlations result in a very slow transition to scaling, regarded as a form of thermodynamic equilibrium. The standard methods of statistical analysis (variance methods) are a source of misleading information in this case: the first being mistaking the regime of transition to scaling as either ordinary or anomalous scaling. The second is that the scaling value, as determined by the evaluation of the second moment, might significantly depart from the correct one. All the VM techniques are shown to be affected by this limitation, while the DEA is the only technique always yielding the correct scaling value, if the scaling condition applies. The application to the study of DNA sequences reported in this paper yields:

1) a striking example of how the standard techniques can produce misleading conclusions

2) a suggestive example of the power of the DEA, which, in this case, is able to indicate clearly that both coding and non-coding sequences generate Lévy statistics in the long-time limit.

The DEA is not only a method of scaling detection. Its entropic nature gives also useful insights into the regime of transition from dynamics to thermodynamics. It is possible to prove that in the special case where the time series is generated by fluctuations around a locally varying bias the regime of transition to the scaling regime is significantly delayed. The ordinary techniques of analysis mistaken this transition regime as a form of anomalous memory, while the DEA makes it possible to establish the genuine nature of the process under study. This is left as a subject for further applications.

[1] National Center for Biotechnology Information. http://www.ncbi.nlm.nih.gov/

[2] C.-K. Peng, S.V. Buldyrev, A.L. Goldberger, S. Havlin, F. Sciortino, M. Simons and H.E. Stanley, Nature **356**, 168 (1992).

[3] W. Li, Int. J. Bifurcation Chaos Appl. Sci. Eng. 2, 137 (1992); W. Li and K. Kaneko, Europhys. Lett. **17**, 655 (1992); W.Li, T. Marr, and K. Kaneko, Physica (Amsterdam) D **75**, 392 (1994).

[4] R.F. Voss, Phys. Rev. Lett. **68**, 3805 (1992).

[5] S.V. Buldyrev, A.L. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H.E. Stanley, Phys. Rev. E **47**, 4514 (1993).

[6] A.K. Mohanty and A.V.S.S. Narayana Rao, Phys. Rev. Lett. 84, 1832 (2000).

[7] B. Audit, C. Thermes, C. Vaillant, Y. d'Aubenton-Carafa, J.F. Muzy, and A. Arneodo, Phys. Rev. Lett. **86**, 2471 (2001).

[8] P. Allegrini, M. Barbi, P. Grigolini and B.J. West, Phys. Rev. E **52**, 5281 (1995).

[9] P. Allegrini, M. Buiatti, P. Grigolini and B.J. West, Phys. Rev. E **57**, 4558 (1998).

[10] P. Allegrini, M. Buiatti, P. Grigolini and B.J. West, Phys. Rev. E **58**, 3640 (1998).

[11] P. Allegrini, P. Grigolini, B.J. West, Phys. Rev. E **54**, 4760 (1996).

[12] C. -K. Peng, S.V. Buldyrev, S. Havlin, M. Simons, H.E. Stanley and A.L. Goldberger. Phys. Rev. E **49**, 1685 (1994).

[13] A. Arneodo, E. Bacry, P.V. Graves, and J.F. Muzy, Phys. Rev. Lett. **74**, 3293 (1995).

[14] N. Scafetta, P. Hamilton, P. Grigolini, Fractals **9**, 193 (2001).

[15] A.I. Khinchin *Mathematical Foundations of Statistical Mechanics*, Dover, New Yoor (1949).

[16] Y. Bar-Yam, *Dynamics of Complex Systems* (Addison-Wesley, Reading Mass, 1997).

[17] B. B. Mandelbrot, *Fractal Geometry of Nature*, W.H. Freeman Co. (1988).

[18] B.V. Gnedenko and A.N. Kolmogorov, Limit Distributions for Sums of Independent Random Variables, Addison-Wesley Publishing Company, Inc. Cambridge (1954).

[19] R. Metzler and J. Klafter, Phys. Rep. **339**, 1 (2000).

[20] H. E. Hurst, R. P. Black, Y. M. Simaika, *LongTerm Storage: An Experimental Study* (Constable, London).

[21] J. Feders, *Fractals*, Plenum Publishers, New York (1988).

[22] C.E. Shannon, W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, Urbana (1963).

[23] G. Tréfan, E. Floriani, B.J. West, and P. Grigolini, Phys. Rev. E **50**, 2564 (1994).

[24] M. Bianucci, R. Mannella, B.J. West, and P. Grigolini, Phys. Rev. E **51**, 3002 (1995).

[25] G.M. Zaslavsky, *Physics of Chaos in Hamiltonian Systems*, Imperial College Press, London (1998).

[26] T. Horita, H. Hata, R. Ishizaki, and H. Mori, Ptrog. Theor. Phys. **83**, 1065 (1990);R. Ishizaki, H. Hata, T. Horita, and H. Mori, Prog. Theor. Phys. **84**, 179 (1990).

[27] J. Klafter, G. Zumofen, and M.F. Shlesinger, Europhys. Lett. **25**, 565 (1994)

[28] L. Bonci, P. Grigolini, A. Laux and R. Roncaglia, Phys. Rev. A **54**, 112 (1996); M. Stefancich, P. Allegrini, L. Bonci, P. Grigolini, and B. J. West, Phys. Rev. E **57**, 6625 (1998).

[29] T. Geisel, J. Nierwetberg, and A. Zacherl, Phys. Rev. Lett. **54**, 616 (1985).

[30] M. Annunziato, P. Grigolini, Phys. Lett. A **269**, 31 (2000).

[31] G. Bramanti, P. Grigolini, M. Ignaccolo, G. Raffaelli, work in progress.

[32] M. Bologna, P. Grigolini, B.J. West, submitted to Chem. Phys.

[33] G. Zumofen, J. Klafter and M.F. Shlesinger, Springer Lecture Notes in Physics **519**, 15 (1998).

[34] J. Klafter and G. Zumofen, J. Phys. Chem. **98**, 7366 (1994).

[35] M. Bologna, P. Grigolini, and J. Riccardi, Phys. Rev. E, **60**, 6435 (1999).

[36] G. Zumofen, J. Klafter, Phys. Rev. E 47, 851 (1993).

[37] D. Bedeaux, K. Lakatos, and K. Shuler, J. Math. Phys. **12**, 2116 (1971).

[38] L. E. Reichl, *Statistical Physics*, J. Wiley. New York (1998).

[39] M. Buiatti, P. Grigolini, L. Palatella, Physica A **268**, 214 (1999).

[40] A. Torcini, R. Livi, A. Politi, A dynamical approach to protein folding, Journ. of Biol. Phys. 27 (2): 181-203 2001

[41] M. Araujo, S. Havlin, G.H. Weiss, and H.E. Stanley, Phys. Rev. A **43**, 5207 (1991).

[42] R. Mannella, P. Grigolini, and B.J. West, Fractals 2, 81 (1994)

[43] V. Seshadri, B.J. West, Proc. Natl. Acad. Sci. USA 79 (1982) 4501.

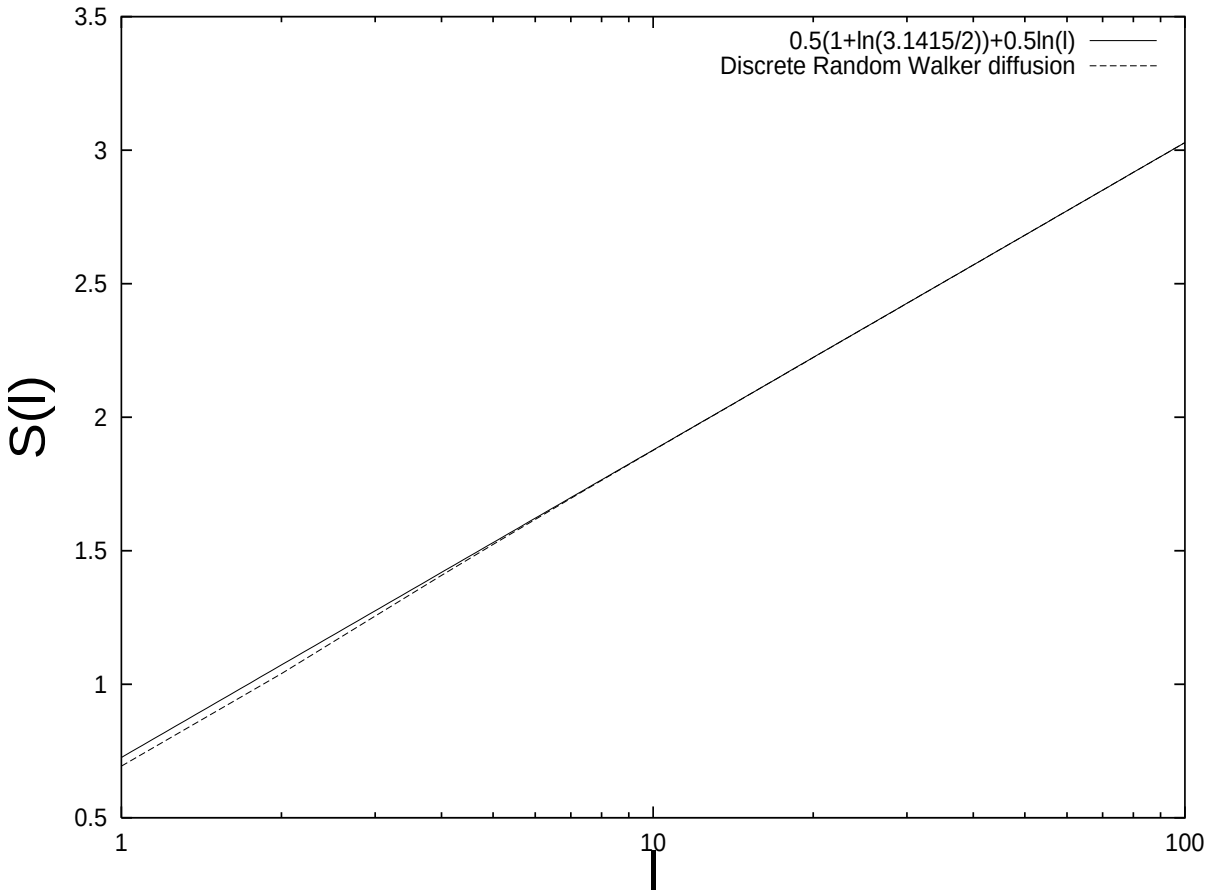


FIG. 1. Diffusion entropy of a random walker as a function of the number of jumps l. The dashed lines and the solid line denote the discrete diffusion entropy $S_d(l)$, of Eq.(35) and the continuous prescription of Eq.(5), respectively. After a short transient the dashed line converges to the solid line.

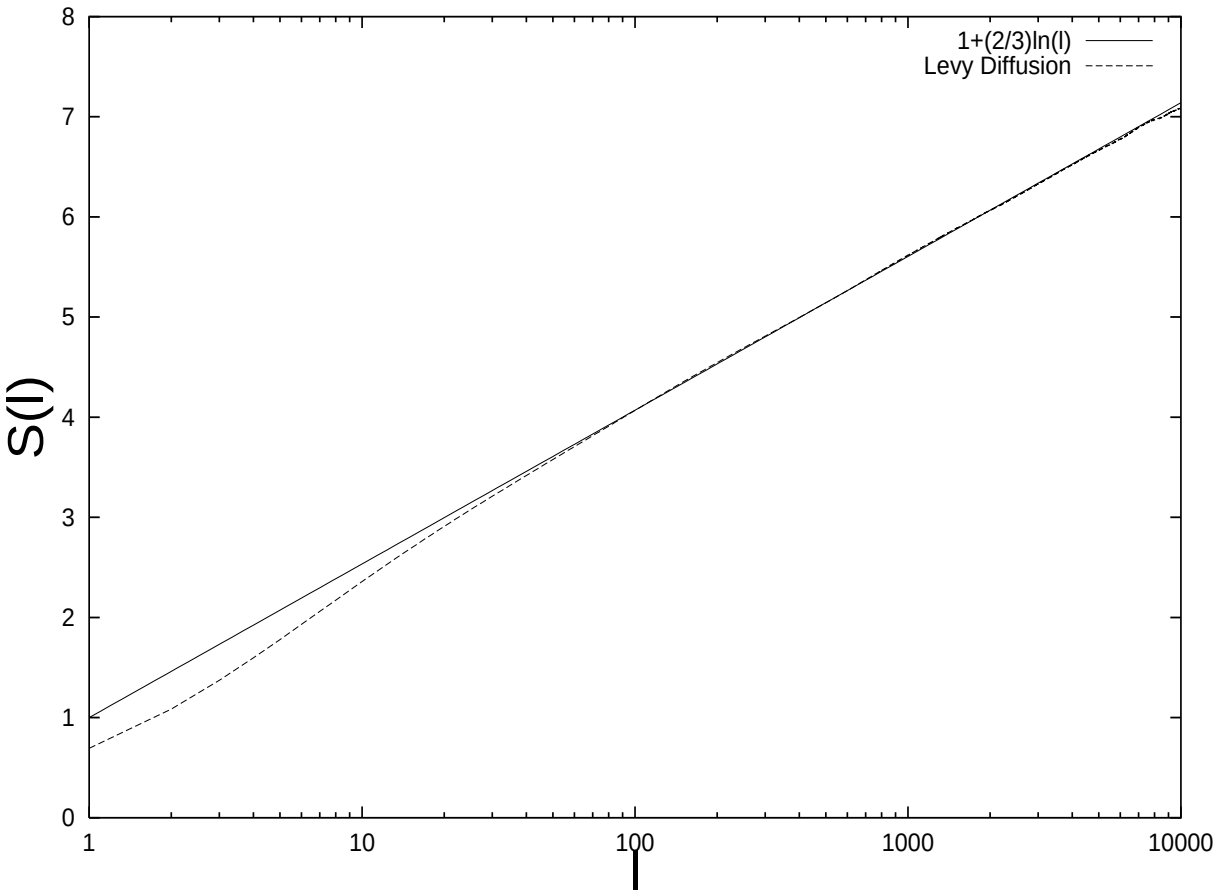
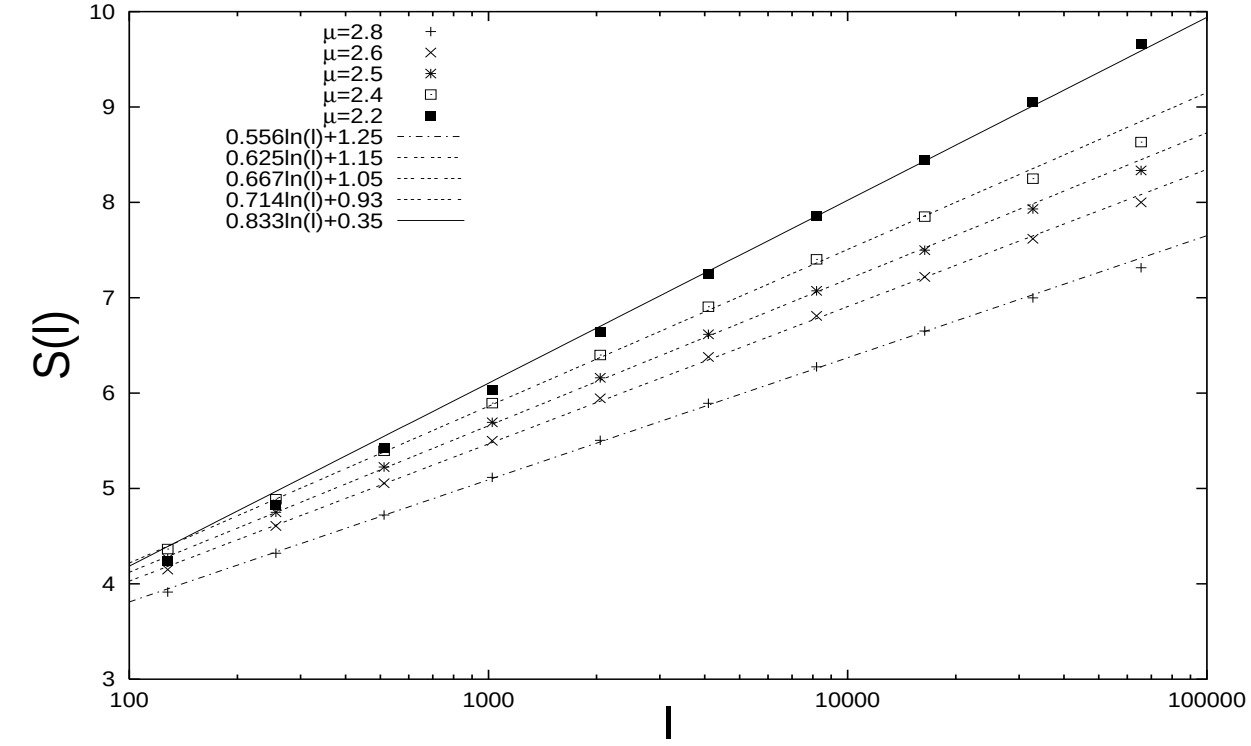


FIG. 2. Diffusion entropy of the Lévy process generated by an articial sequence, ξ_i, corresponding to the power coefficient $\mu = 2.5$ and $T = 0$. The dashed line is the diffusion entropy, $S_d(l)$, in the discrete-space perspective given by the Eq.(33). The solid line is the diffusion entropy $S(l)$ in the continuous-space perspective given by the Eq.(39). After an initial transient, the dashed line converges to the solid line.



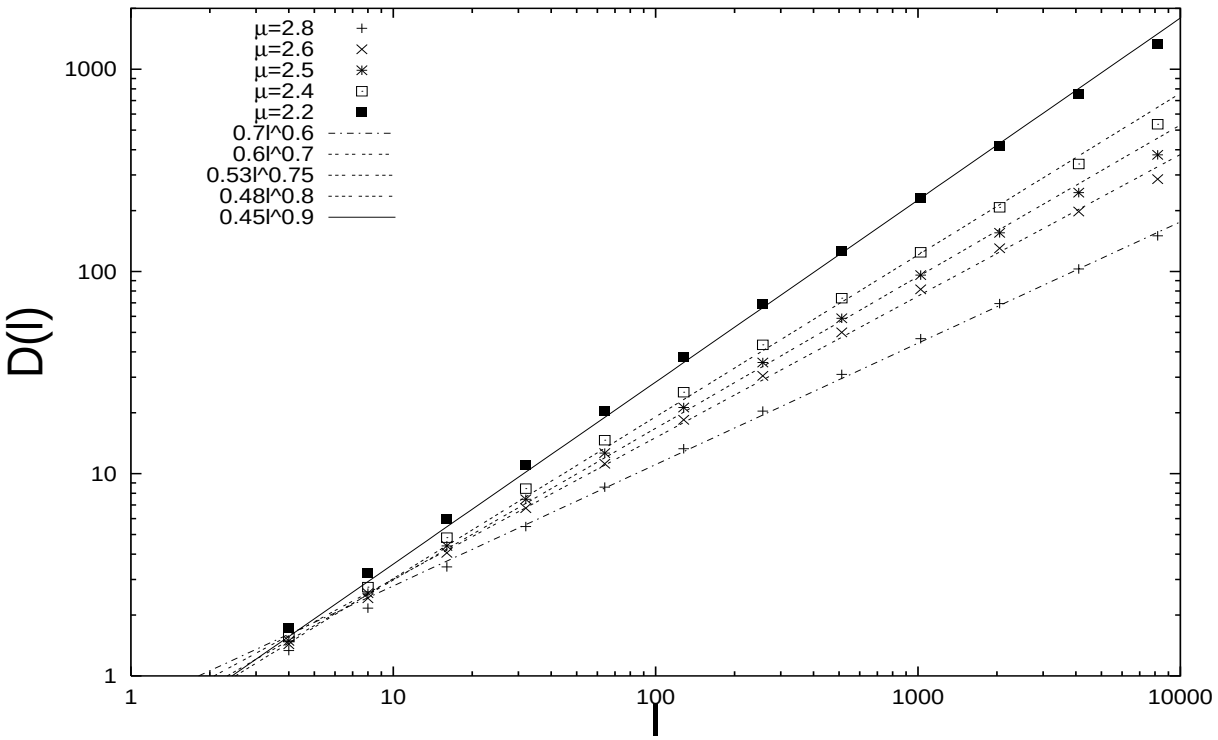


FIG. 3. Diffusion entropy and Standard Deviation of the Lévy process generated by articial sequences ξ_i corresponding respectively to five different values of the power coefficient μ, namely: $\mu = 2.8, 2.6, 2.5, 2.4, 2.2$, and $T = 0$. The numerical results of the Diffusion Entropy Analysis (DEA) (3a) and of the Standard Deviation Analysis (SDA) (3b), reported in symbols, are in perfect agreement with the theoretical predictions, reported as fitting lines and obtained by using respectively the values of the pdf scaling exponent δ and of the esponents H in Table I.

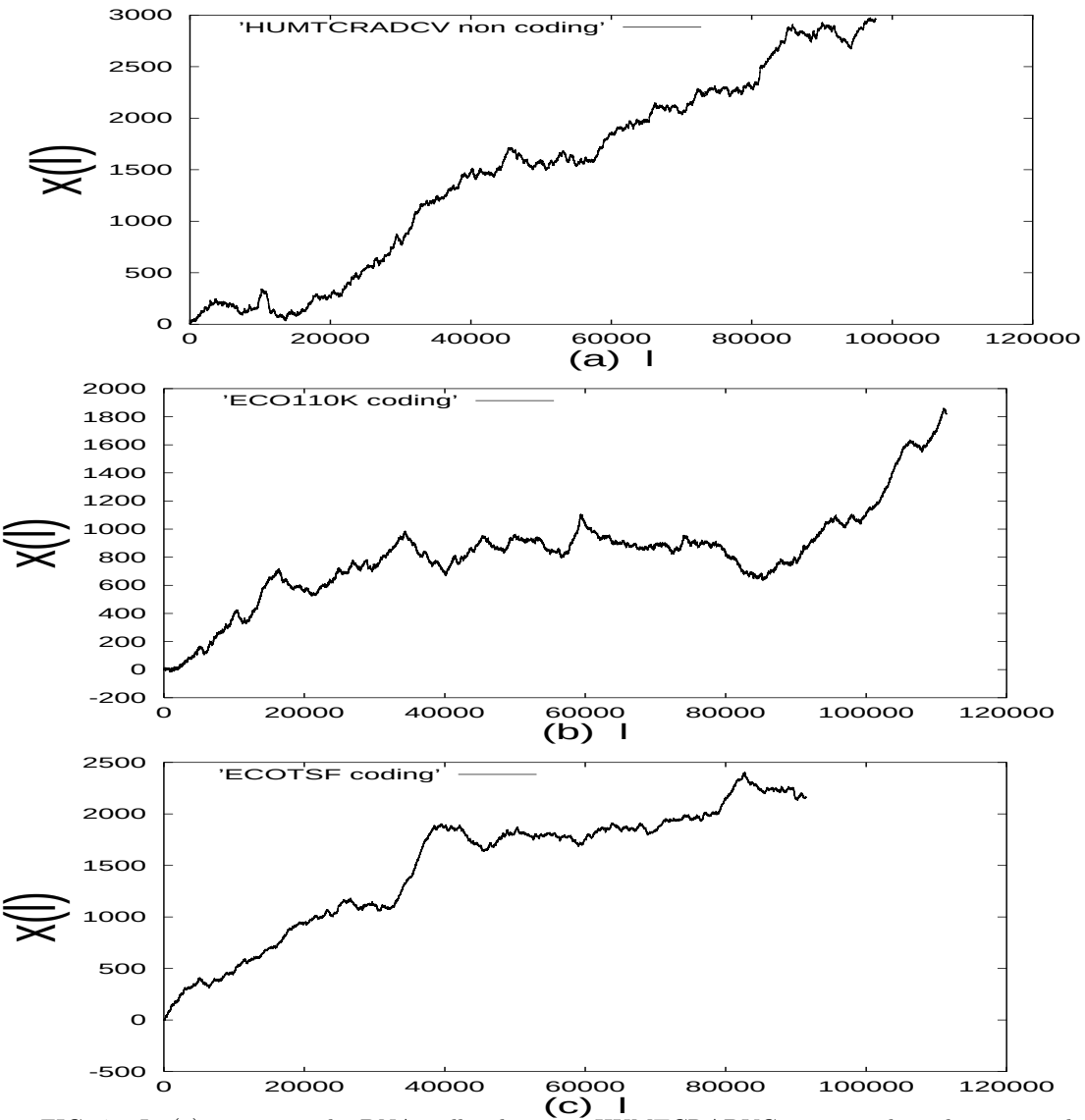


FIG. 4. In (a) we report the DNA walk relative to HUMTCRADVC, a non-coding chromosomal fragment. In (b) and (c), we report the DNA walk relative to ECO110K and ECOTSF, two coding genomic fragments.

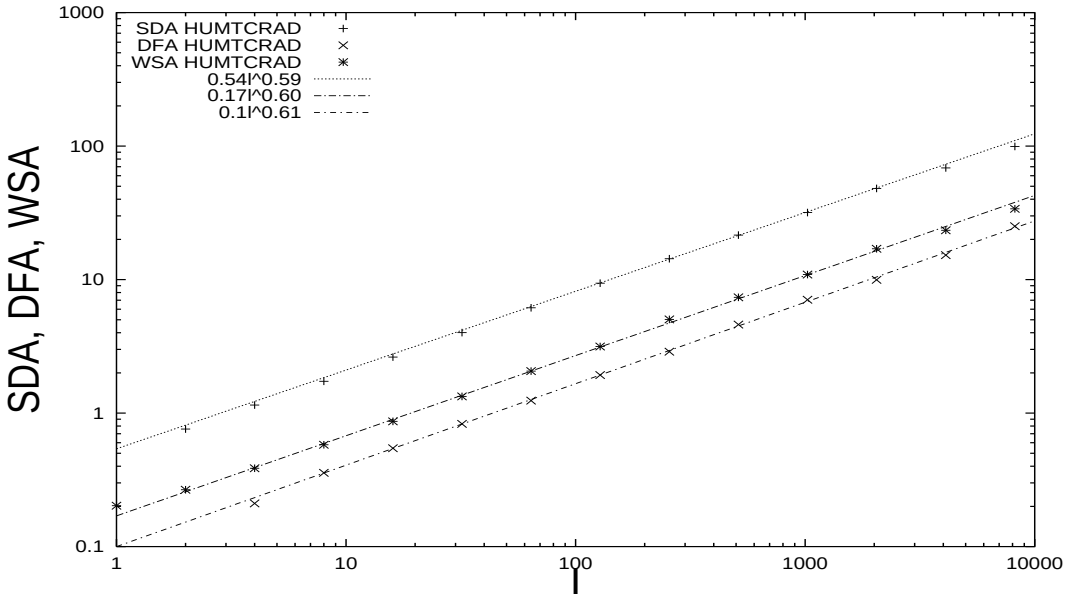


FIG. 5. Application of the three variance methods SDA, DFA and WSA to the sequence HUMTCRADVC, the non-coding chromosomal fragment. The three methods give the same exponent H. In fact we get $H = 0.59 \pm 0.01$ (SDA), $H = 0.60 \pm 0.01$ (DFA) and $H = 0.61 \pm 0.01$ (WSA), where the differences are within the error bars. Moreover H is the same both at short-time and long-time regions (i.e. $H' = H$).

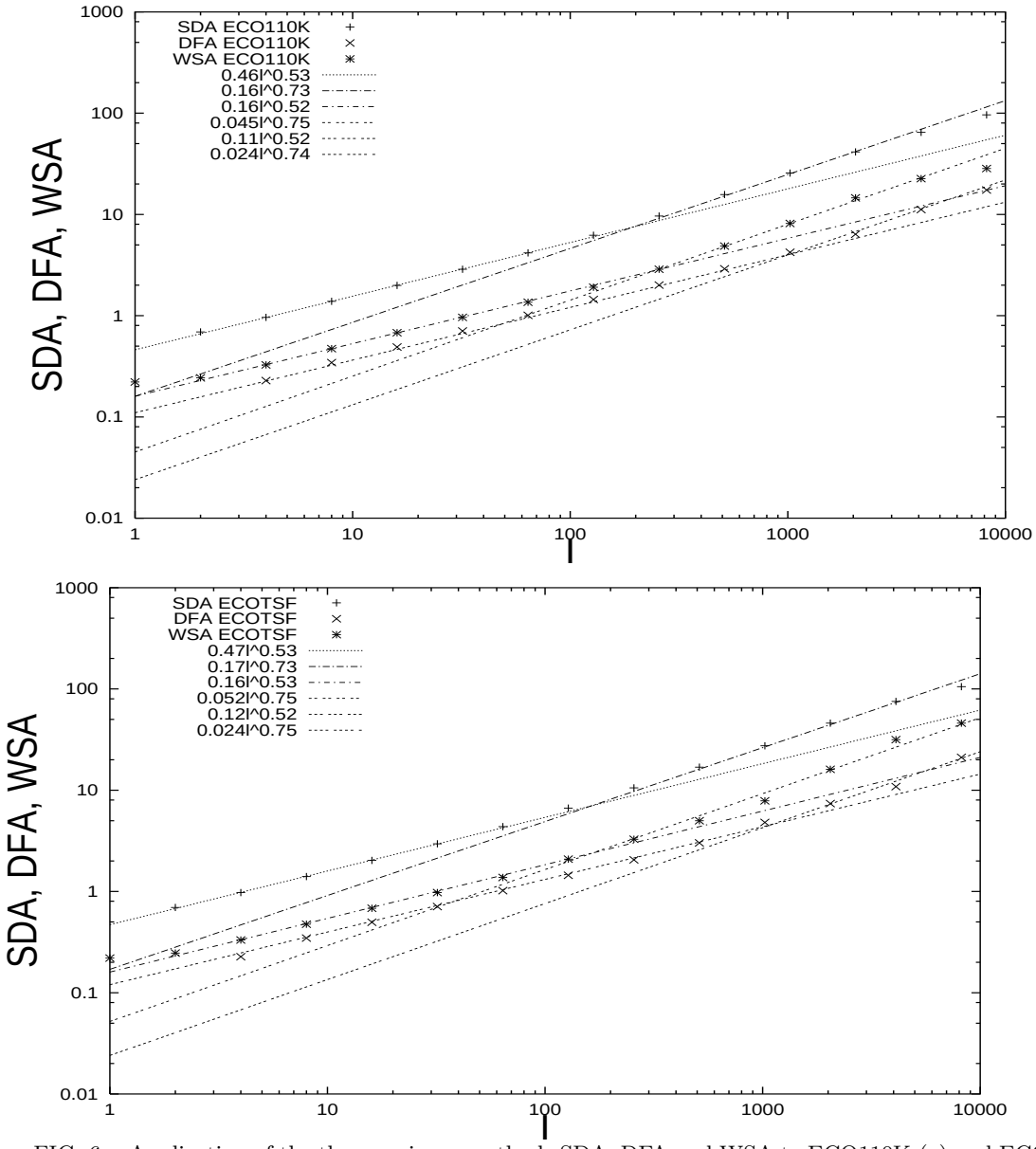


FIG. 6. Application of the three variance methods SDA, DFA and WSA to ECO110K (a) and ECOTSF (b), the two coding genomic fragments. The scaling exponent H' in the short-time region is 0.53 ± 0.01 (SDSA), 0.52 ± 0.01 (DFA), 0.52 ± 0.01 (WSA). The scaling exponent H in the long-time regions is 0.73 ± 0.01 (SDSA), 0.75 ± 0.01 (DFA), 0.74 ± 0.01 (WSA).

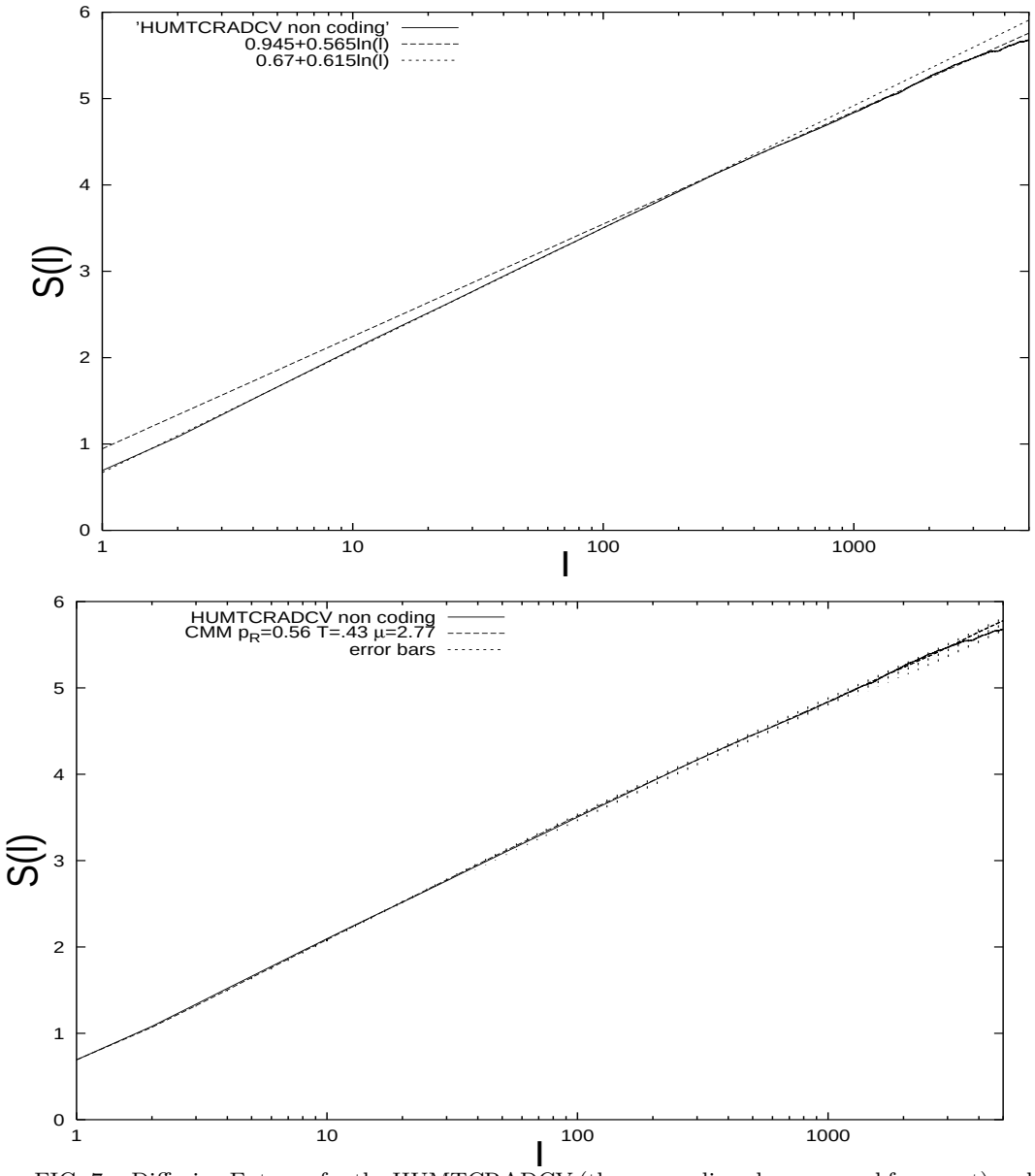


FIG. 7. Diffusion Entropy for the HUMTCRADCV (the non-coding chromosomal fragment) and its CMM simulation. Fig.7a shows that the DE analysis results in a scaling changing with time. The slope of the two straight lines is $\delta' = 0.615 \pm 0.01$ in the short-time region, and $\delta = 0.565 \pm 0.01$ in the long-time regime. Fig.7b shows the comparison between the DEA of the real non-coding sequence and an artificial sequence corresponding to the CMM model: $p_R = 0.56 \pm 0.02$, $T = 0.43$, $\mu = 2.77 \pm 0.02$.

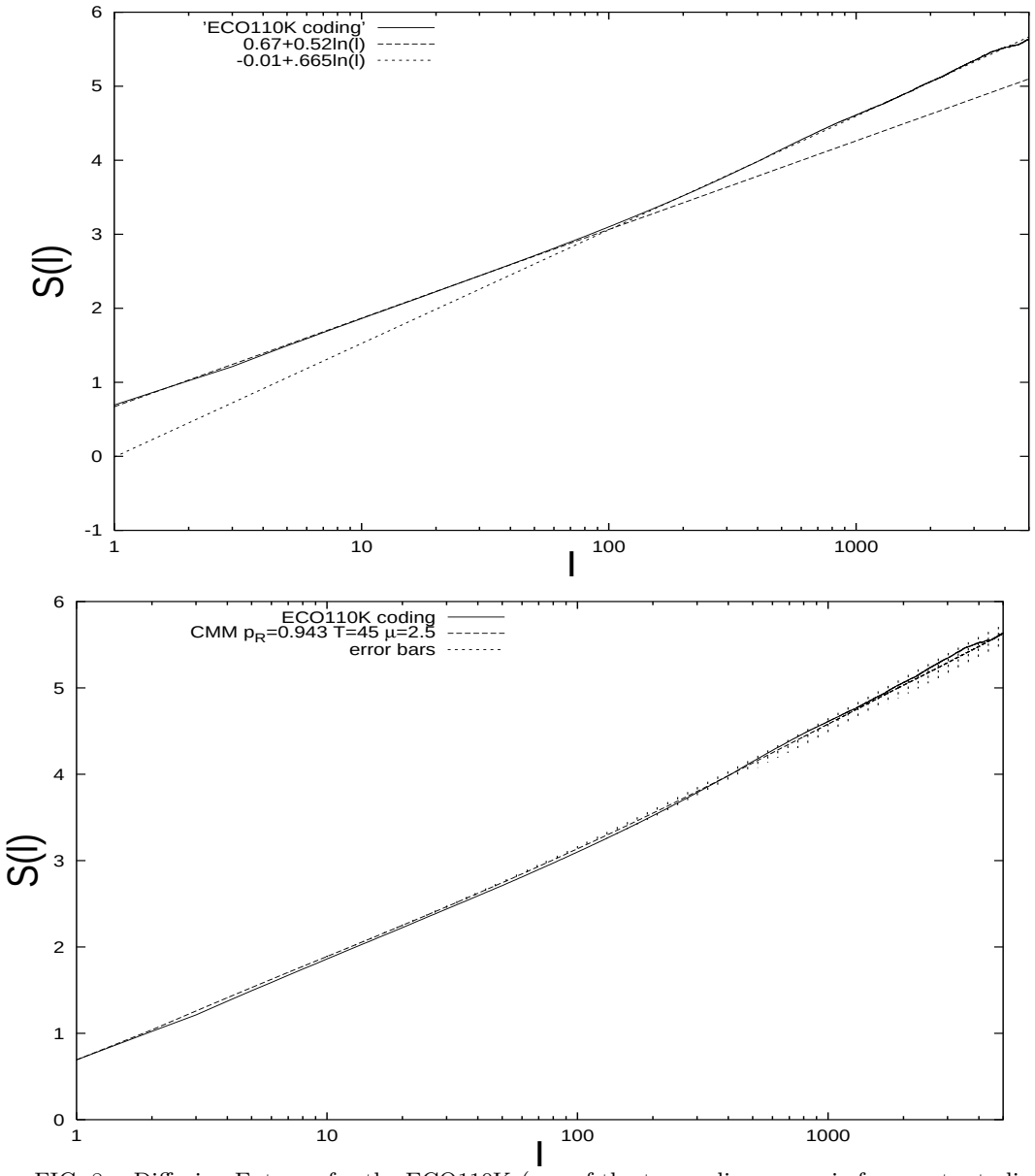


FIG. 8. Diffusion Entropy for the ECO110K (one of the two coding genomic fragments studied) and its CMM simulation. Fig.8a shows that the DEA results in a scaling changing with time. The slope of the two straight lines is $\delta' = 0.52 \pm 0.01$ at short-time regime, and $\delta = 0.665 \pm 0.01$ at long-time regime. Fig.8b shows the comparison between the DE analysis of the real coding sequence and an artificial sequence corresponding to the CMM model: $p_R = 0.943 \pm 0.01$, $T = 45$, $\mu = 2.5 \pm 0.02$.

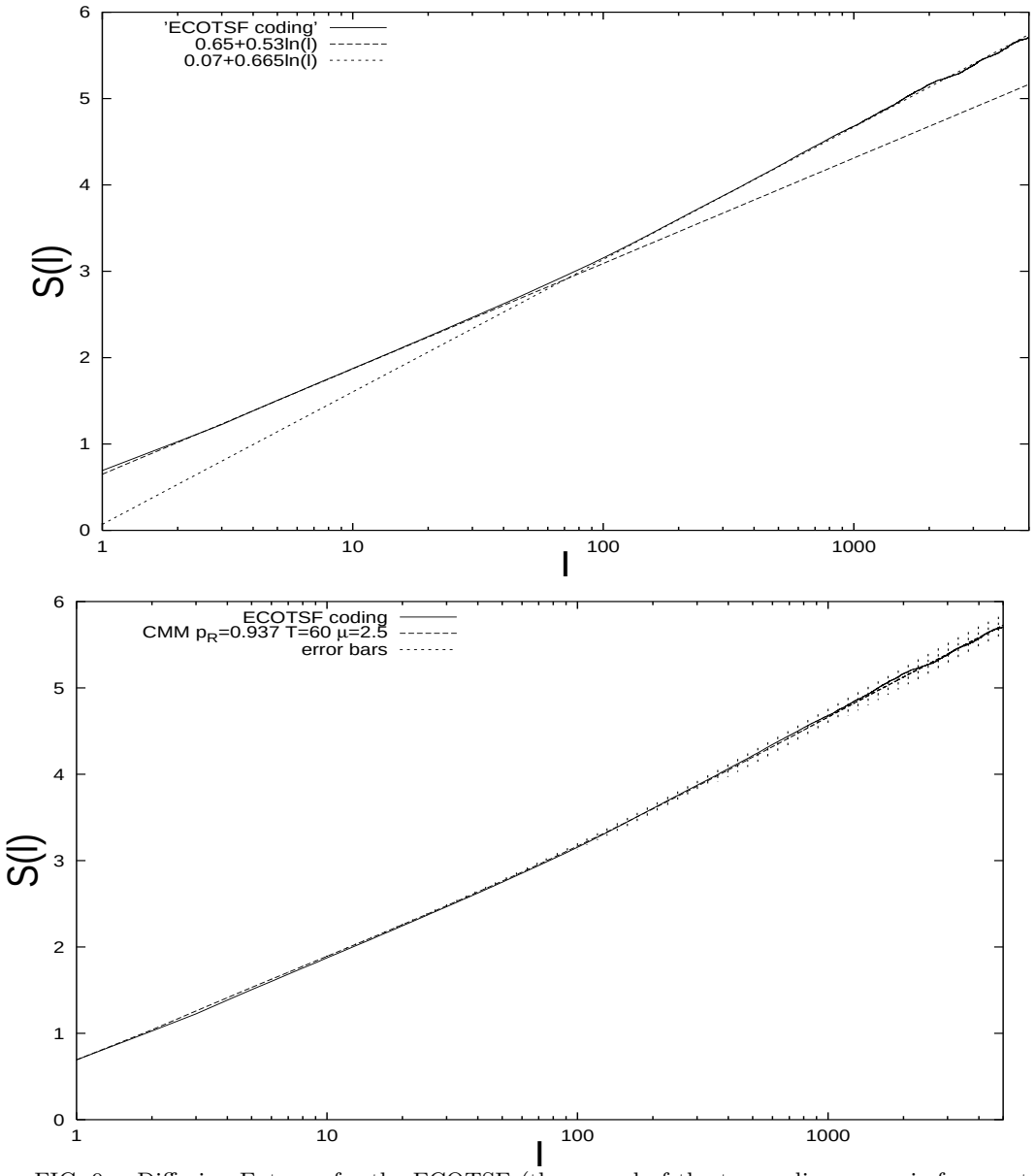


FIG. 9. Diffusion Entropy for the ECOTSF (the second of the two coding genomic fragments studied in this paper) and its CMM simulation. Fig.9a shows that the DEA results in a scaling changing with time. The slope of the two straight lines is $\delta' = 0.53 \pm 0.01$ in the short-time region, and $\delta = 0.665 \pm 0.01$ in the long-time regime. Fig.9b shows the comparison between the DEA of the real coding sequence and an artificial sequence corresponding to the CMM model: $p_R = 0.937 \pm 0.01$, $T = 60$, $\mu = 2.5 \pm 0.02$.

μ	2.200	2.400	2.500	2.600	2.800
H	0.900	0.800	0.750	0.700	0.600
δ	0.833	0.714	0.667	0.625	0.556

TABLE I. In the first line from the top we report the power indices of the inverse power law distributions used to create the artificial sequences studied in Fig.3. In the second line from the top we report the corresponding Hurst coefficients, prediction of Eq.(20). In the third line from the top we report the true scaling, namely the Lévy scaling of Eq.(24).

Non-Coding	N	H	δ_H	δ
HUMTCRADCV	97630	0.61	0.56	0.56
CELMYUNC	9000	0.71	0.63	0.635
CHKMYHE	31109	0.78	0.69	0.70
DROMHC	22663	0.72	0.64	0.65
HUMBMYHZ	28437	0.58	0.54	0.54
Coding				
ECO110K	111401	0.74	0.66	0.66
ECOTSF	91430	0.74	0.66	0.66
LAMCG	48502	0.85	0.77	0.76
CHKMYHN	7003	0.74	0.66	0.66
DDIMYHC	6680	0.68	0.61	0.61
DROMYONMA	6338	0.69	0.62	0.64
HUMBMYH7CD	6008	0.63	0.57	0.58
HUMDYS	13957	0.69	0.62	0.62

TABLE II. Values of the scaling exponents H and δ for a set of different coding and non-coding sequences. In the first column we report the GenBank name of the sequence [1], and in the second column the length N of the sequence. For all measures the error is ± 0.01. δ_H in the fourth column is the theoretical value for δ if the Lévy ondition applies, Eq.(41). If the length of the genome is larger than 20,000 the fitted region is $100 < l < 2000$. If the length of the genome is shorter than 20,000, the statistics are not very good for large l. In this case, the fitted region is $20 < l < 200$.